

CLAIMS IN CURRENT FORM
U.S. Patent Application Serial No. 10/661,245

1. (previously presented) A method implemented in a computer system, for clustering a string, the string including a plurality of characters, the method including:

identifying R unique n-grams $T_{1\dots R}$ in the string;

for every unique n-gram T_S :

if the frequency of T_S in a set of n-gram statistics is not greater than a first threshold:

clustering the string with a cluster associated with T_S ;

otherwise:

for every other n-gram T_V in the string $T_{1\dots R, \text{ except } S}$:

concluding that the frequency of n-gram T_V is greater than the first threshold, and in response:

if the frequency of n-gram pair T_S-T_V is not greater than a

second threshold:

clustering the string with a cluster associated with the n-gram pair T_S-T_V ;

otherwise:

for every other n-gram T_X in the string $T_{1\dots R, \text{ except } S \text{ and } V}$:

clustering the string with a cluster associated with the n-gram triple $T_S-T_V-T_X$;

where $T_{1\dots R}$ is a set of n-grams, R is the number of elements in

$T_{1\dots R}$, and T_S , T_V , and T_X are members of $T_{1\dots R}$, and S, V, and X are integer indexes to identify members of $T_{1\dots R}$.

2. (original) The method of claim 1 further including compiling n-gram statistics.
3. (original) The method of claim 1 further including compiling n-gram pair statistics.
4. (canceled)
5. (canceled)
6. (previously presented) A method implemented in a computer system, for clustering a string, the string including a plurality of characters, the method including:
 - identifying R unique n-grams $T_{1\dots R}$ in the string;
 - for every unique n-gram T_S :
 - if the frequency of T_S in a set of n-gram statistics is not greater than a first threshold:
 - clustering the string with a cluster associated with T_S ;
 - otherwise:
 - for $i = 1$ to Y :
 - for every unique set of i n-grams T_U in the string $T_{1\dots R}$, except S :
 - if the frequency of the n-gram set $T_S - T_U$ is not greater than a second threshold:
 - clustering the string with a cluster associated with the n-gram set $T_S - T_U$;
 - if the string has not been associated with a cluster with this value of T_S :
 - for every unique set of $Y+1$ n-grams T_{UY} in the string $T_{1\dots R}$, except S :

clustering the string with a cluster associated with the
Y+2 n-gram group T_S-T_{UY} ,
where $T_{1...R}$ is a set of n-grams, R is the number of elements in
 $T_{1...R}$, and T_S , T_V , and T_X are members of $T_{1...R}$, and S, V,
and X are integer indexes to identify members of $T_{1...R}$.

7. (original) The method of claim 6 where Y = 1.
8. (original) The method of claim 6 further including compiling n-gram statistics.
9. (original) The method of claim 6 further including compiling n-gram group statistics.
10. (canceled)
11. (canceled)
12. (canceled)